Data Analysis of OpenStack Nova Project.

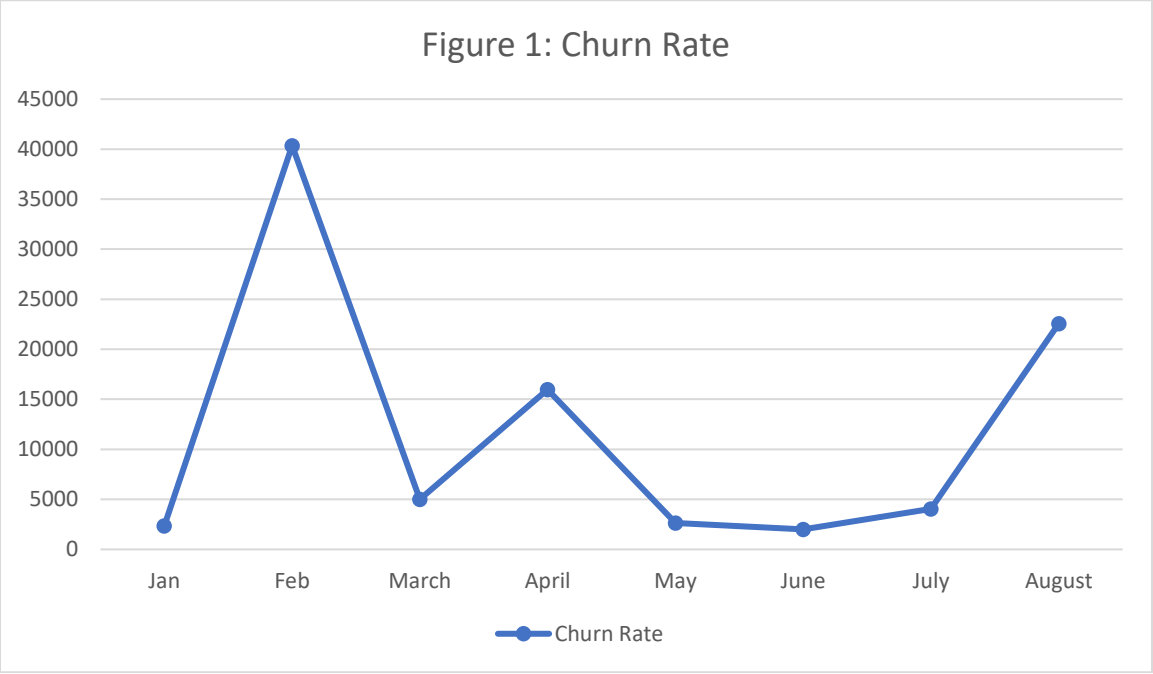
Approach of Getting Data

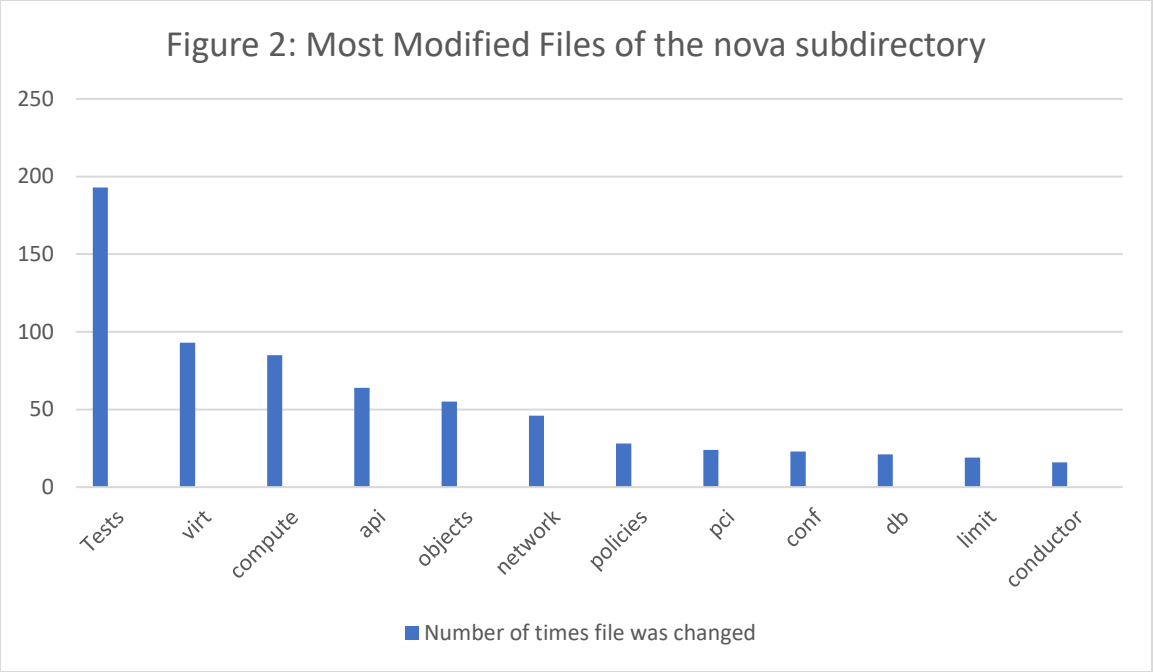
I decided to use the method of page scraping to get all the commits the nova project. The GitHub API was not sufficient to analyze large amounts of data at once because there is a very small limit on how many calls I can make to the API within a short time. I used the python package called BeautifulSoup to get the html of the GitHub pages (commit pages, nova subdirectory page) that I needed. After getting those pages, I was able to analyze them to retrieve the exact data needed for this analysis.

Observations

Upon analysis of the changes made to the files of the nova subdirectory, my observation is that most of the changes made in the past six months, were done on the test folder. As shown in Figure 2, the number of changes to the test folder is about double the changes of the next file in the rank.

I checked the monthly Churn rate of the project and the average churn rate per month is 11,865. From the data shown in Figure 1, it seems the general trend is one month of a lot of work and then one month downtime. This variance is evident from the months of January to May.





Conclusion

From the data shown, it seems like a lot of the work done on the nova project over the past 6 months was a lot of testing. This may not be entirely true though since I was only analyzing the master branch; maybe analyzing the commits made to the other branches may give a different result on most edited files and it may also produce a different monthly churn pattern.